

Cloud Computing

Virtual Resources and Distributed Cloud Applications

Chapter 7: Energy-Efficient and Carbon-Aware Computing



Distributed and Operating Systems
Electrical Engineering and Computer Science
Technische Universität Berlin

Overview

- **Carbon Footprint of Computing**
- Energy-Efficient Computing
 - Energy Demand of Hardware
 - Energy-Proportional Computing
 - Power Usage Effectiveness
- Demand Response
- Carbon-Aware Computing
 - Carbon-Accounting and the GHG-Protocol
 - Optimizing for Low Carbon Intensity
 - Optimizing for Renewable Excess Energy

Carbon Footprint of Computing

- Data centers account for 1 - 2% of global energy consumption and around 2.7% within EU
- Including the supply chain and life cycle of hardware, computing accounts for 2.3 - 3.9 % of global carbon emissions
- While algorithms and hardware became more efficient over time, the overall energy demand of cloud computing keeps increasing

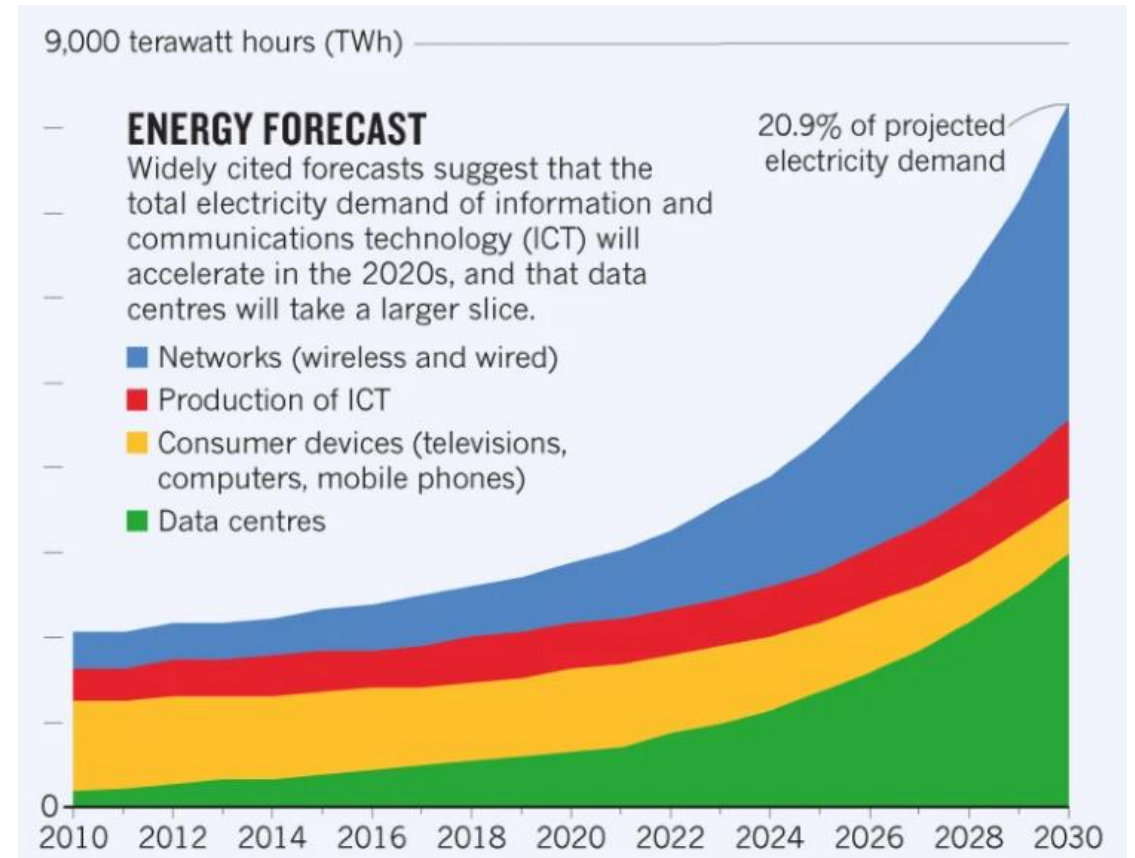


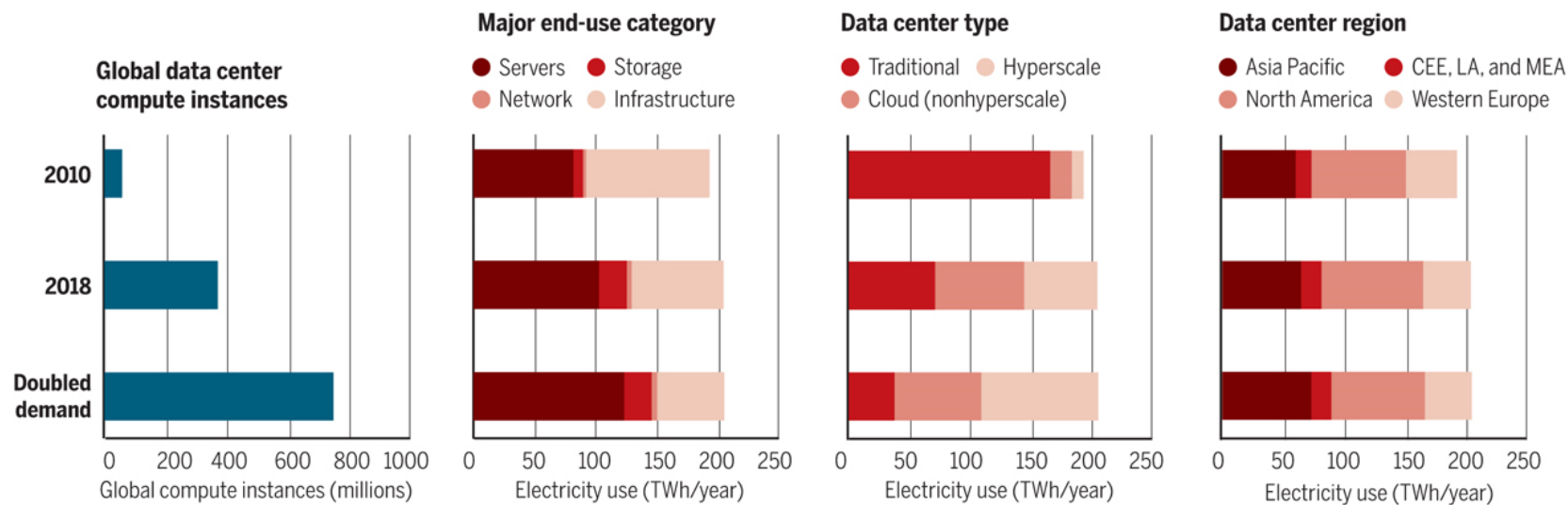
Figure from (2)

Carbon Footprint of Computing

- A continuous shift to highly optimized hyperscale data centers has kept the energy demand of cloud computing fairly stable in the past
- This is expected to change, as efficiency gains become harder to achieve

Historical energy usage and projected energy usage under doubled computing demand

Doubled demand (relative to 2018) reflects current efficiency trends continuing alongside predicted growth in compute instances.



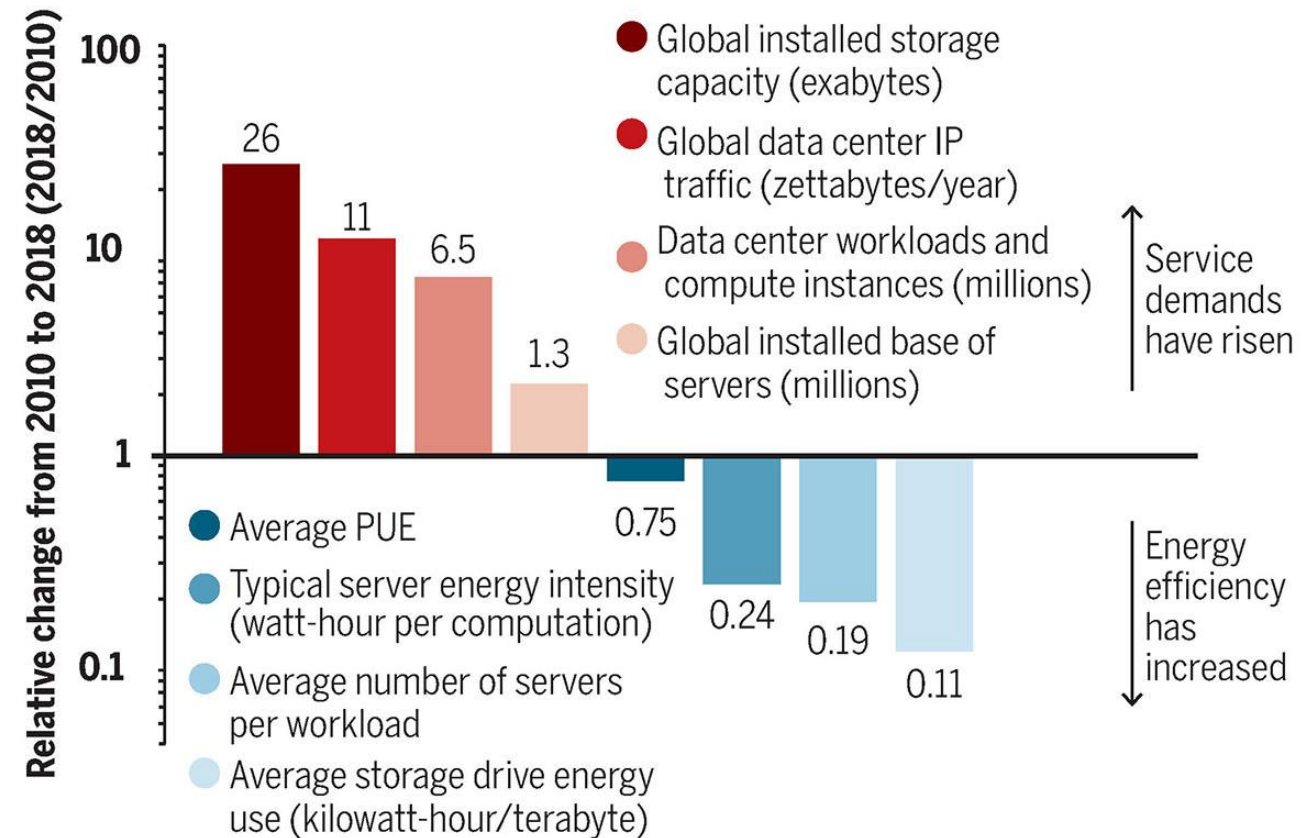
CEE, LA, and MEA, Central and Eastern Europe, Latin America, and Middle East and Africa; TWh, terrawatt-hour.

Figure from (3)

Carbon Footprint of Computing

- **Rebound effects** describe the phenomenon where increases in efficiency ramp up consumption
- e.g. more energy-efficient GPUs lead to the raise of deep learning, which is one of the most energy-intensive compute applications today

Trends in global data center energy-use drivers



PUE, power usage effectiveness; IP, internet protocol.

Figure from (3)

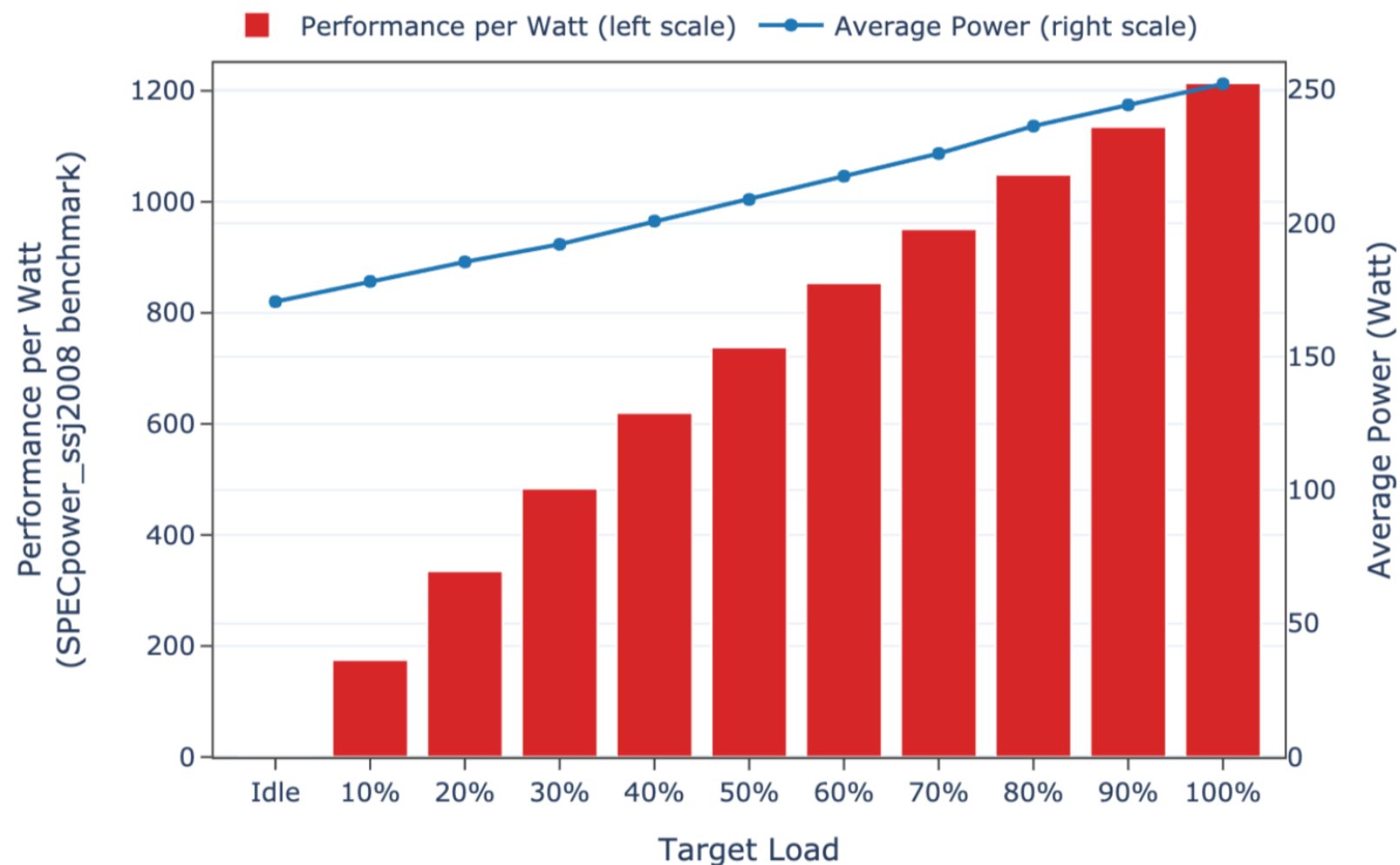
Energy Demand of Hardware

Recap: Power (W) x Time = Energy (1 Joule = 1 Ws; 3600 J = 1 Wh)

- When describing the energy demand of hardware, we differentiate between
 - **Static power consumption:** This power is consumed as soon as the device is turned on (and idle)
 - **Dynamic power consumption:** This power demand scales with the load on the device
- Most devices follow a relatively linear power profile
- But how big is their fraction of static power?

Performance per Watt

- Amount of computation performed per unit of consumed energy
- Example from the *SPECpower_ssj2008* power benchmark
- The performance (i.e. FLOPS) per Watt increases with higher utilization

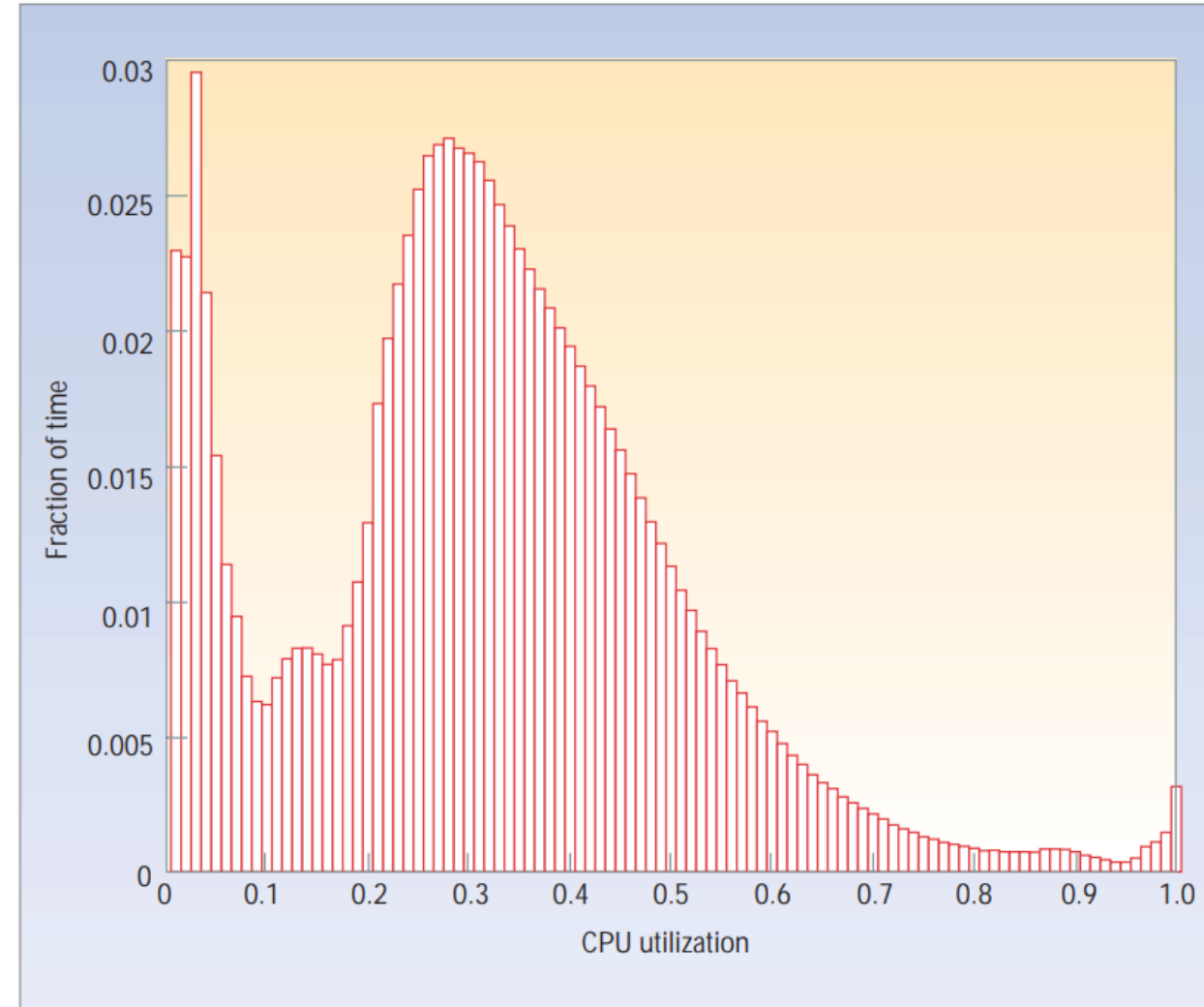


Energy-Proportional Computing

- Cloud data centers are never fully utilized to offer scalability
- To meet throughput and latency service-level agreements, cloud applications are not designed to fully utilize their VM either

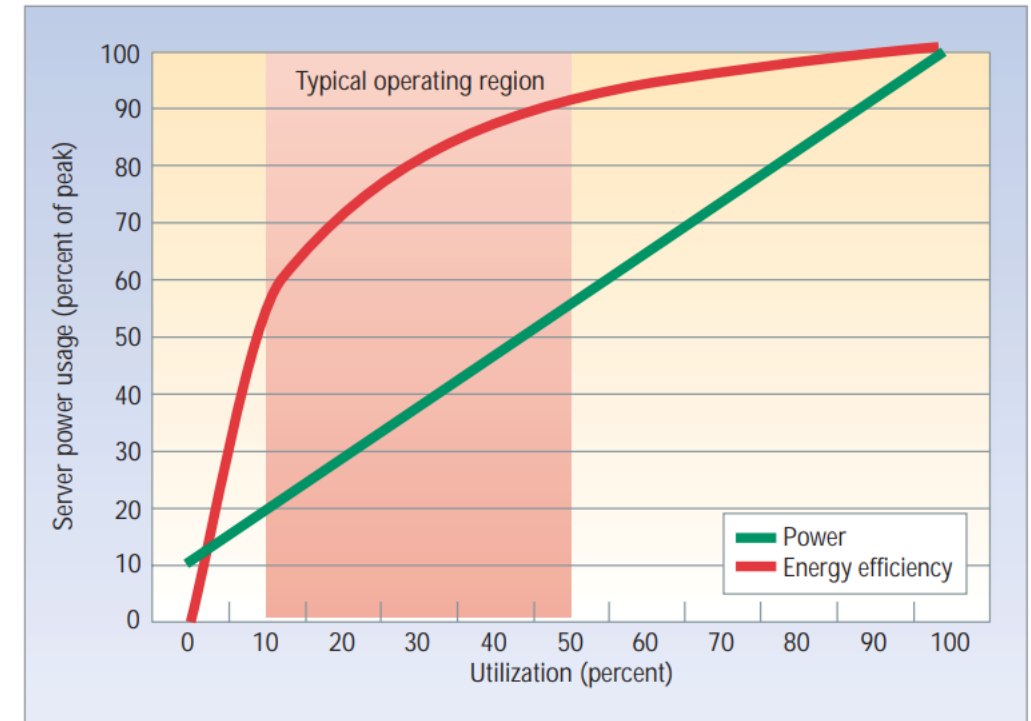
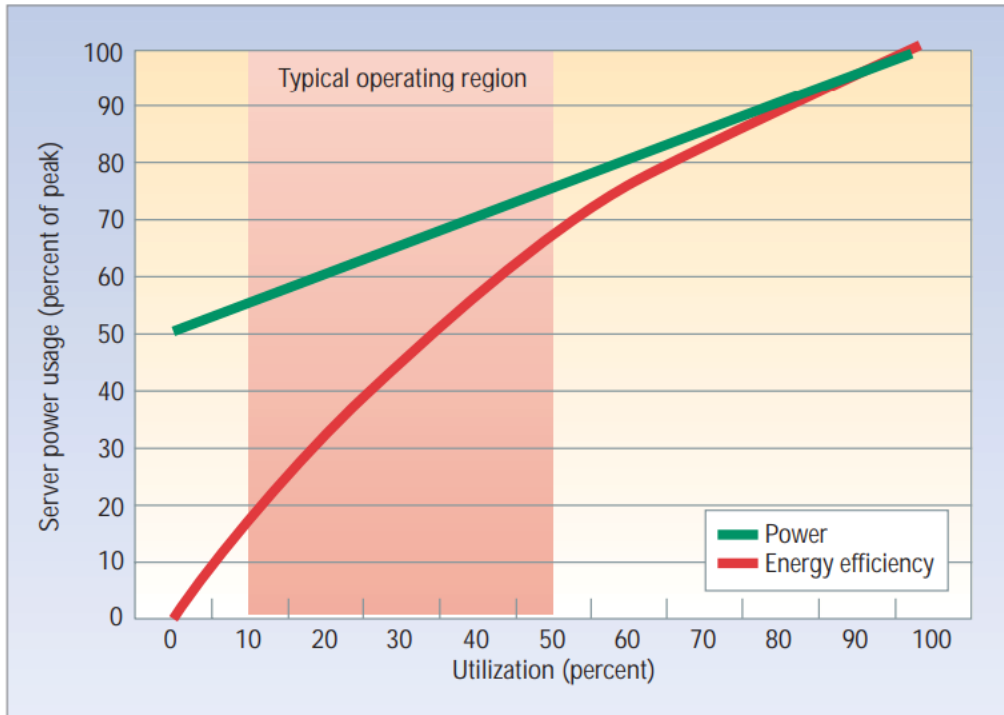
Average CPU utilization of
>5000 Google servers, see (3)

Most servers operate between
10% and 50% of their
maximum utilization levels



Energy-Proportional Computing

- Non-optimized data center
 - e.g. HPC cluster which is expected to run close to 100% at all times
- More energy-proportional data center
 - e.g. cloud data center where the expected utilization is 10-50%



To derive energy efficiency, we simply divide utilization by its corresponding power value

Energy-Proportional Computing

- Energy-proportional data centers require optimizations on many levels
- Energy-proportional hardware
 - CPUs, GPUs, ...
 - memory and storage
 - inter-data center network
- Energy-aware scheduling
 - Should consolidate workload on as few servers as possible, so the remaining servers can be turned off
- Energy-aware data center design
 - E.g. data centers should be designed so that parts of the cooling system or uninterruptible power supply can be shut down together with servers

Power Usage Effectiveness

- Power Usage Effectiveness (PUE) describes how efficiently a data center uses energy.

$$\frac{\text{all electricity used by a data center}}{\text{electricity used by computing hardware}}$$

- An ideal PUE is 1.0 → All electricity is solely used by computing hardware
- In 2022 the average PUE was 1.55, while highly-optimized hyperscale data centers can achieve values of around 1.1
- Meaning that only an additional 10% are required for overhead such as cooling, lights, pumps, ventilation, etc.

Demand Response

- Electric grids **always need to be balanced** in terms of supply and demand
 - This is increasingly hard with variable renewable energy sources like solar and wind
 - In the future, we need more flexible consumers
- In **demand response**, power consumers adapt their electricity usage to better match demand with supply
 - e.g. major electricity consumers like steel plants, refrigerated warehouses, etc. can buy electricity at reduced rates for offering demand response
 - Increasingly, data centers join demand response programs to lower costs

Data Center Demand Response

- Opportunities
 - Data centers often execute flexible workloads
 - Some workloads can be executed at a later point in time
 - Other workloads can be migrated to another data center in a different region
- Challenges
 - Most current data centers are not power proportional
 - Cloud operators often have no knowledge about which workloads are flexible
 - Data centers must be reliable
 - ◆ short-term demand response notifications endanger SLAs
 - ◆ We need to forecast demand response periods to plan schedules in advance

Demand Response Example From Google



<https://cloud.google.com/blog/products/infrastructure/using-demand-response-to-reduce-data-center-power-consumption>

Carbon Accounting

- Quantify the emissions caused by an organization or service
- So far mainly on a voluntary basis or to improve public perception
- Around the world, legislation is slowly increasing regulation and/or introducing carbon pricing mechanisms
 - e.g. UK's "Net Zero Strategy" by 2050
- Likewise, companies define carbon reduction plans
 - e.g. Google's "24/7 Carbon-Free Energy" by 2030
 - Introducing novel renewable energy certificates that incentivize to consume energy directly **when** and **where** it is produced

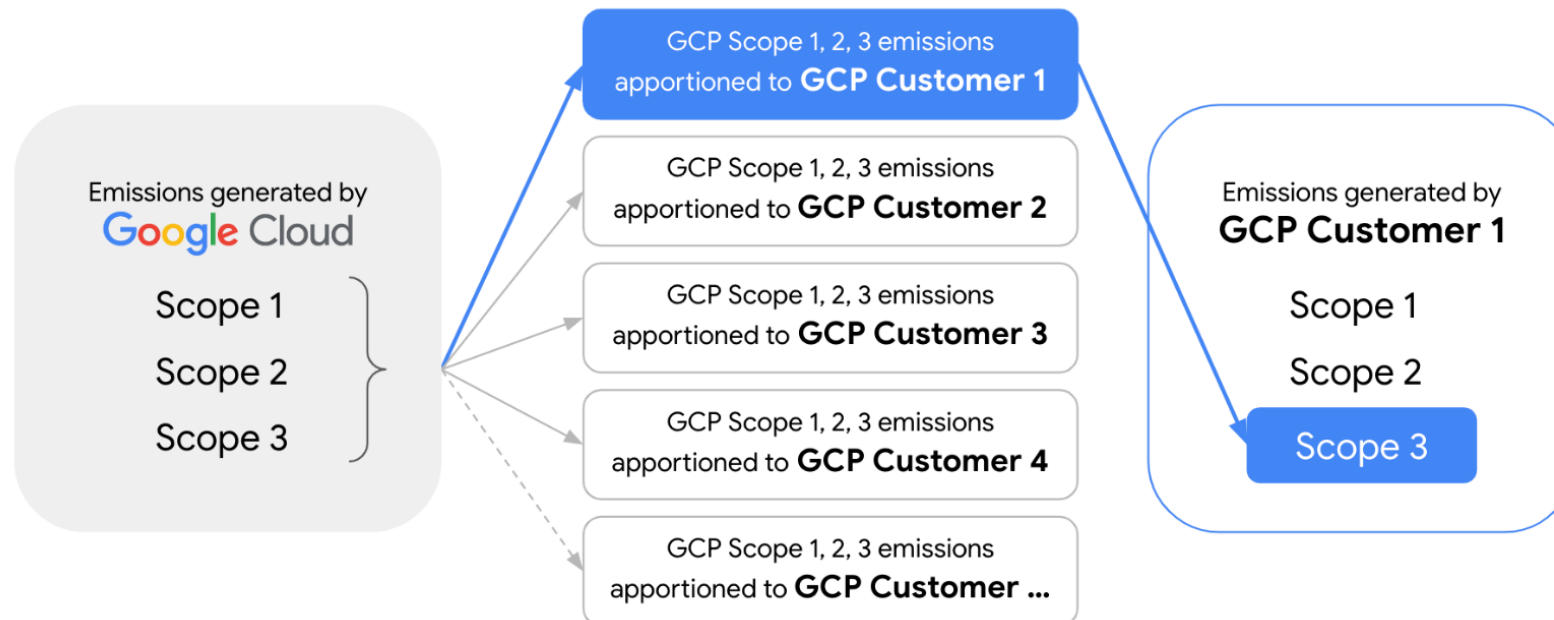
Greenhouse Gas Protocol

- The **Greenhouse Gas (GHG) Protocol** is the most widely used greenhouse gas accounting standard
 - **Scope 1** encompasses emissions that an organization owns or controls directly
 - **Scope 2** emissions are caused indirectly, for example through energy purchases
 - **Scope 3** covers emissions occurring in an organization's value chain
- Most emissions of cloud operators are in **scope 2**
- Cloud customers perceive these emissions within their **scope 3**
- Currently, there are few rules on reporting scope 3 emissions
- Companies can “hide” some of their caused carbon emissions by moving computation to the cloud (thereby reducing their own energy consumption)

Greenhouse Gas Protocol

- The GHG-Protocol is applied by big cloud vendors like AWS, GCP, and Azure
- They increasingly report carbon emissions of utilized services, for customers to be able to report their scope 3 emissions

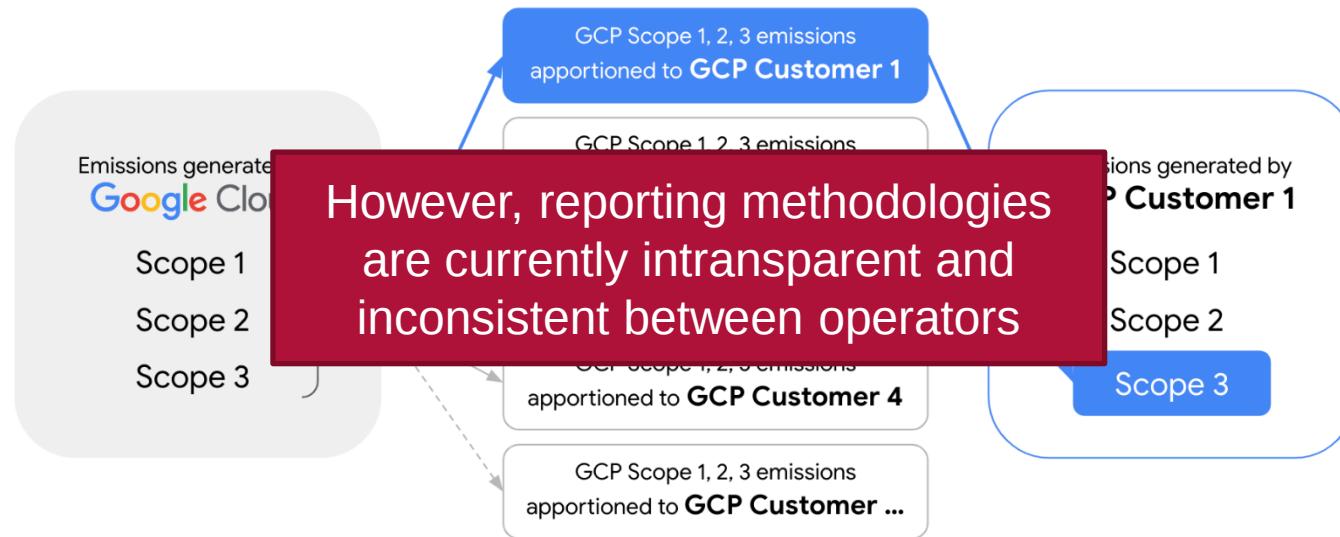
Apportioned GCP's Scope 1, 2, 3 emissions are accounted as part of GCP Customers' Scope 3



Greenhouse Gas Protocol

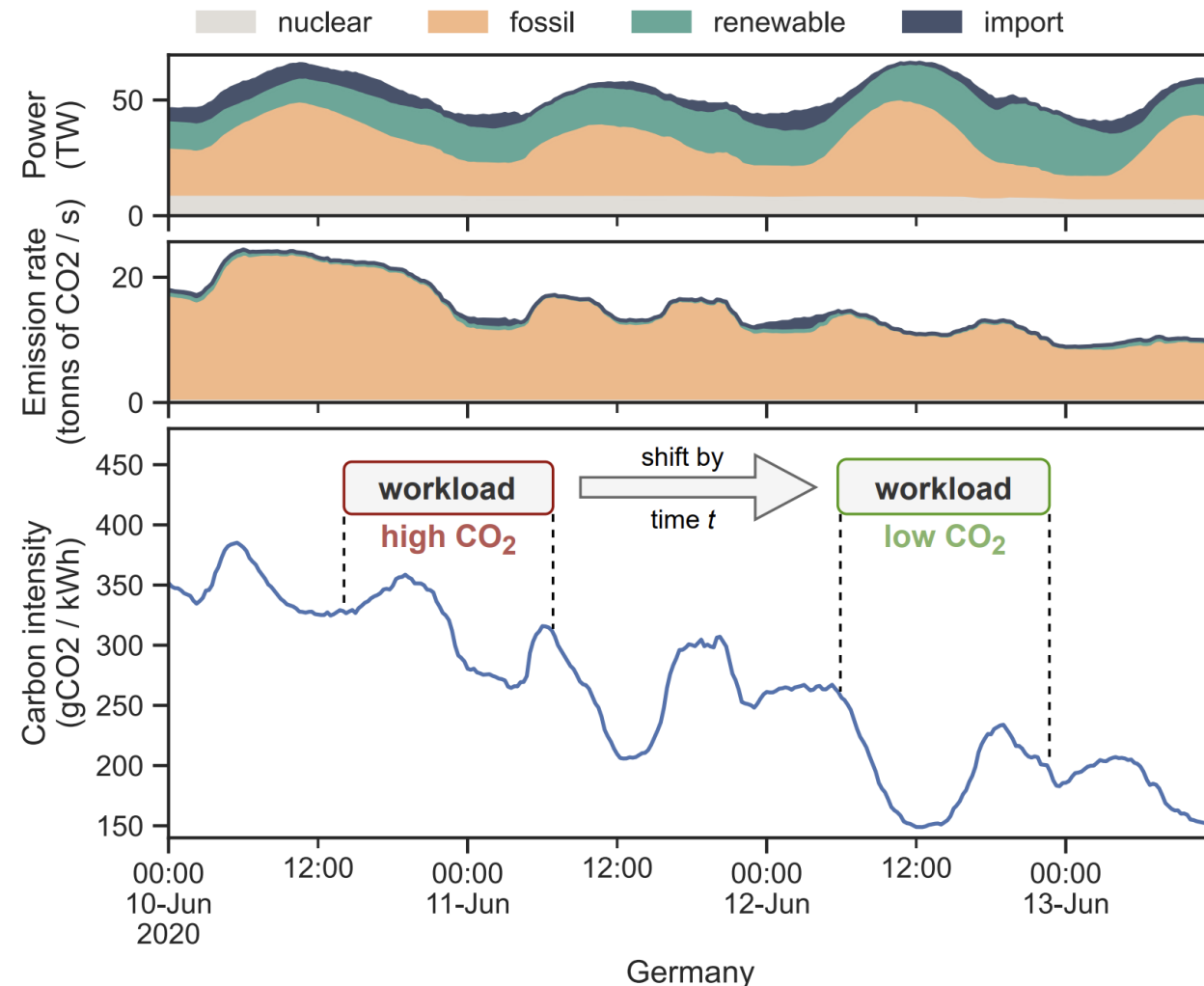
- The GHG-Protocol is applied by big cloud vendors like AWS, GCP, and Azure
- They increasingly report carbon emissions of utilized services, for customers to be able to report their scope 3 emissions

Apportioned GCP's Scope 1, 2, 3 emissions are accounted as part of GCP Customers' Scope 3



Carbon-Aware Computing

- Techniques used for demand response (load shifting & workload migration) can also be applied to reduce operational carbon emissions
- **Carbon-aware computing** aims at aligning the power demand of computing with the availability of low-carbon energy
- Carbon intensity describes the amount of associated CO₂-equivalent emissions when consuming energy



Average Carbon Intensity

- All produced power weighted by the carbon intensity of their power source

Energy Source	Biopower	Solar Energy	Geothermal Energy	Hydropower	Wind Energy	Nuclear Energy	Natural Gas	Oil	Coal
gCO ₂ /kWh	18	46	45	4	12	16	469	840	1001

- Good metric for carbon accounting
 - average carbon intensity is easy to understand and compute
 - power production per source and cross-border flows are publicly reported in the EU though the ENTSO-E and similarly in other parts of the world
 - See <https://app.electricitymaps.com>

Average Carbon Intensity

- Varies between regions, seasons, weekdays, and times of day
- Problem: Reporting areas are very large (e.g. entire Germany) so local curtailments are not captured in this metric

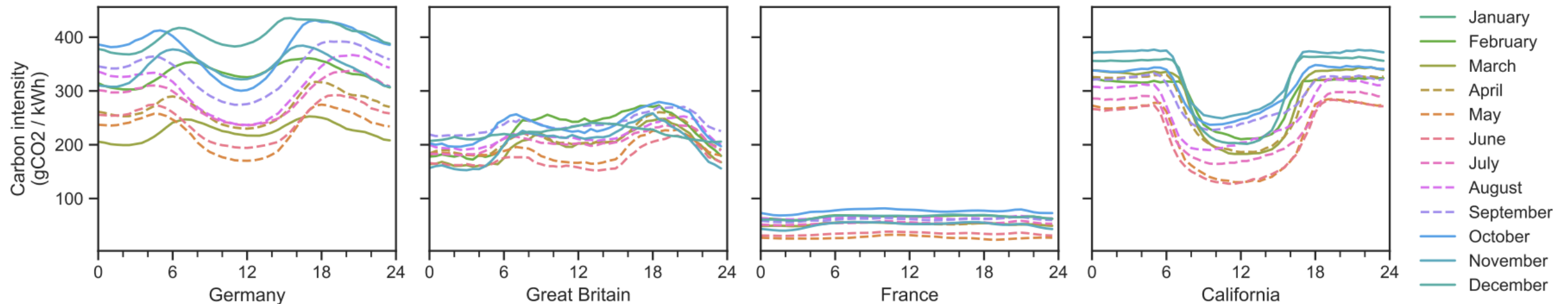
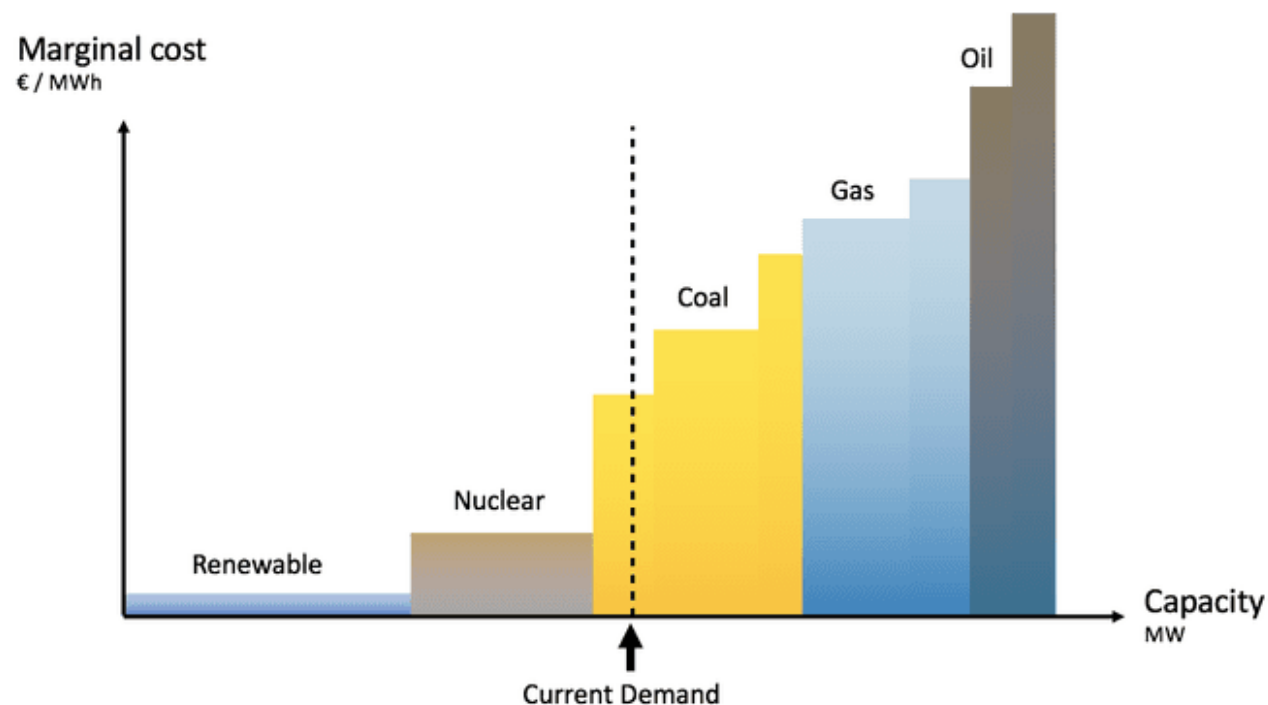


Table and figure from (4)

Marginal Carbon Intensity

- If we consume an **additional** unit of energy, what's its carbon intensity?
 - e.g. at times we have lots of wind and solar in the grid → low average carbon intensity
 - but additional demand must be met by a coal power plant → high marginal carbon intensity
- Good metric for decision making
 - But: In large grids it is very hard to know the marginal power plant
 - We must estimate it, leading to various inconsistent methodologies



<https://www.electricitymaps.com/blog/marginal-emissions-what-they-are-and-when-to-use-them>

Accounting vs. Offsetting

- Two questions with different answers
 - How much am I emitting by consuming electricity at a certain time?
 - What are the emissions avoided if I decide to consume electricity at a certain time?
 - Using more electricity at a given time
 - The first plant to increase its production is the last dispatched one (cheapest plant with spare capacity) named marginal power plant.
 - Example: Grid with 50% wind / 50% gas
 - Marginal emission factor: more consumption = more energy from gas turbine
→ Emissions of ~ 500 gCO₂ for each additional kWh consumed
 - Average emission factor: Grid users get a proportional share of all emissions → footprint is $\sim 250 + x$ (x small) gCO₂/kWh (the average).
- Decision caused 500 gCO₂/kWh but the footprint is $250 + x$ gCO₂/kWh.
- Confusing?

Accounting vs. Offsetting

- Two ways of assigning responsibility
- Average paradigm: responsibility is distributed evenly across grid users, benefits (installing new wind turbine) as well as losses (additional power consumption)
- Marginal (or consequential) paradigm: responsibility based on a causal relationship between decision and consequence.
 - Additional emissions through marginal power plant are directed to the one who caused such a reduction and leaves other users unaffected.
 - Analogously for benefits

Example: Smart Heater

- Heater consuming 500W connected on a grid with 50% wind (@0gCO₂/kWh) and 50% gas (@500 gCO₂/kWh)
 - Average carbon intensity: 250 gCO₂/kWh → Heater emissions $0.5 \times 250 = 125$ gCO₂ / hour
- Scenario 1 (accounting):
 - Turn off the heater for one hour → emissions saved by avoiding production of 500W gas at 500gCO₂ / kWh (marginal footprint) amounts to $0.5 \times 500 = 250$ gCO₂ for that hour.
- Scenario 2 (offsetting):
 - Heater shifts all consumption to an hour where no gas was needed (all electricity from wind) → footprint is now at 0 gCO₂ for that hour
- Result by buying a smart heater:
 - reduced our footprint from 125 g to 0 g (accounting)
 - avoided 250 g of emissions (offsetting).

Renewable Excess Energy

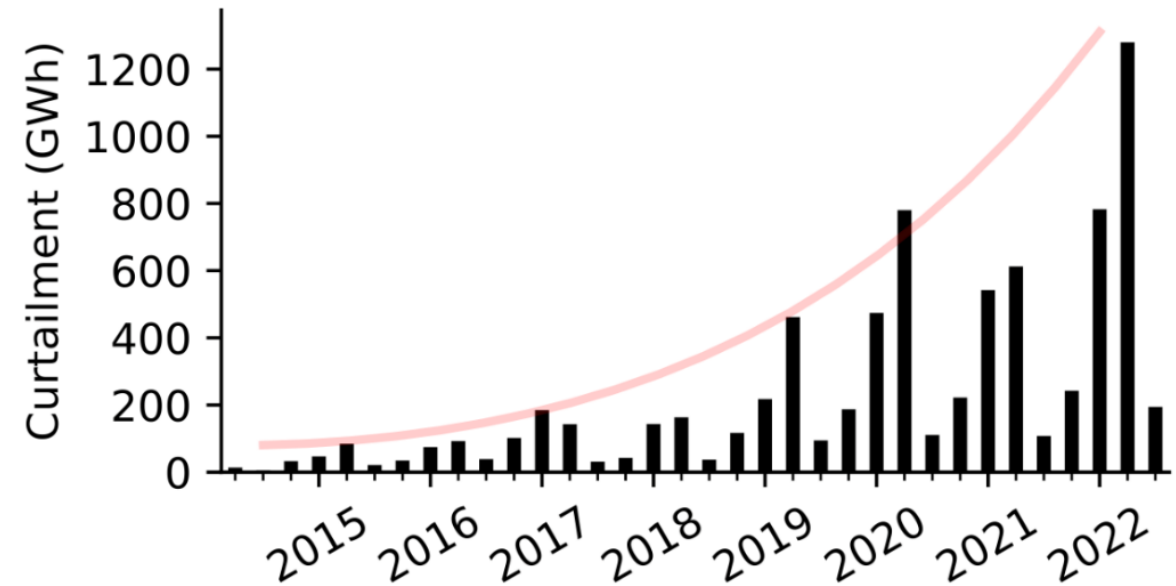
What happens, if we have more power supply than demand?

1. Energy storage (“move” excess energy through time)
 - Expensive and available in limited capacity
 - Conversion losses
 - Charge cycles increase battery aging
2. Transmit to other area (“move” excess energy through space)
 - Neighboring grids often have excess energy at similar times
 - Feeding energy into the public grid offers very low compensation

If neither of these options are available, we must **curtail** excess energy

Renewable Excess Energy

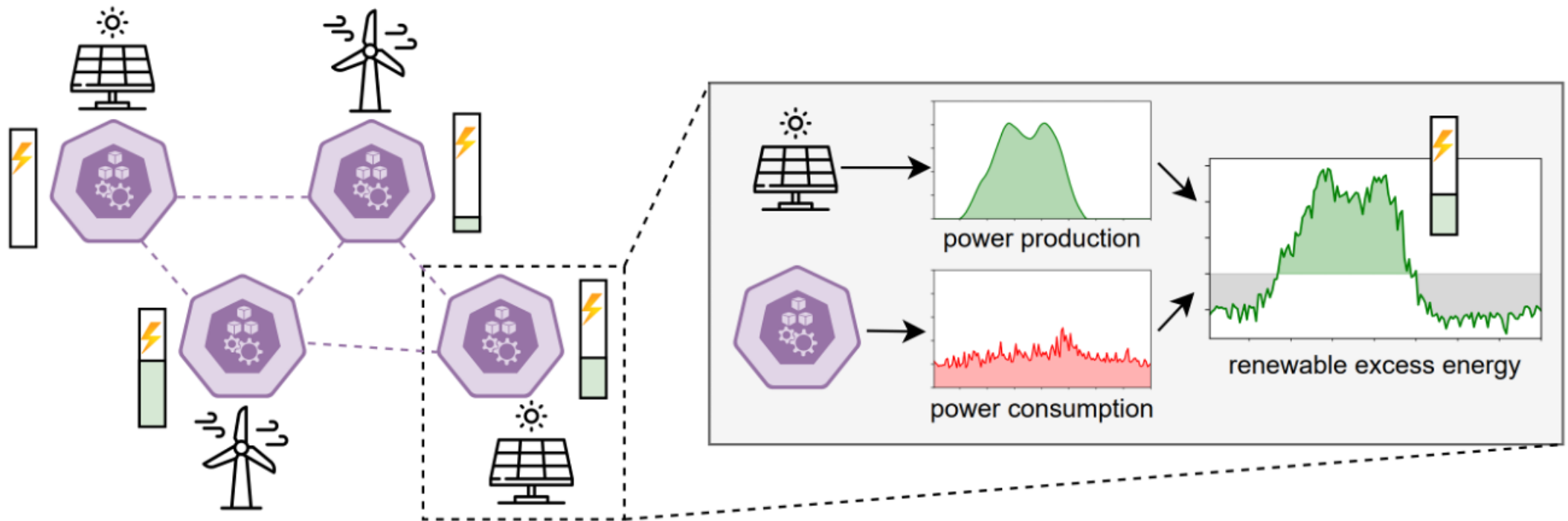
- Curtailment describes the deliberate reduction of power production to balance power supply and demand
- In 2022
 - California curtailed around 7% of their entire solar production
 - Germany lost 5.4 million megawatt-hours of renewable energy due to grid bottlenecks because we fail to transmit power from the coast to the energy-intensive industries in the south



Quarterly wind and solar curtailments by the California ISO.

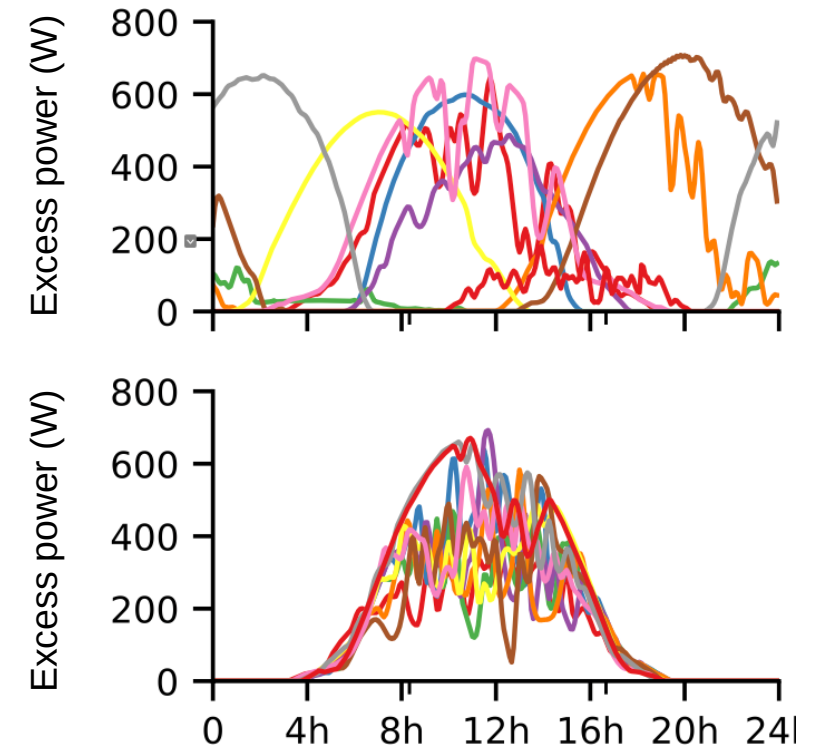
Renewable Excess Energy

- Can we utilize renewable excess energy in data centers by scheduling flexible workloads on space capacity, or even by overclocking fully utilized hardware?



Challenges in Carbon-Aware Scheduling

- It is non-trivial to define meaningful metrics for carbon accounting and optimization
- We heavily rely on the quality of power production forecasts
- Energy efficiency (using less energy) and carbon efficiency (using clean energy) do not always go hand-in-hand
 - e.g. it costs energy to migrate workloads to a region with cleaner power
- How are future computing and energy systems exchanging information in a secure manner?



Sustainable Computing at DOS

If you want to write your **thesis** on topics related to sustainable computing (or are interested in a **student worker** or **PhD position**), we are currently researching on:

- How can we make distributed computing more energy- and carbon-efficient? (4)
- How can carbon-aware scheduling reduce the emissions of ML? (12)
- How can ML be leveraged in carbon-aware scheduling?
- What are realistic testing environments for carbon-aware systems? (13)
 - Have a look at our simulation framework Vessim: <https://github.com/dos-group/vessim>

For more information, please reach out to Philipp Wiesner (wiesner@tu-berlin.de)

References

- (1) Eric Masanet, Arman Shehabi, Nuo Lei, Sarah Josephine Smith, and Jonathan Koomey. “Recalibrating global data center energy-use estimates.” *Science* 367, 2020
- (2) Nicola Jones. “How to stop data centres from gobbling up the world’s electricity.” *Nature* 561, 2018
- (3) Luiz André Barroso and Urs Hölzle. “The Case for Energy-Proportional Computing.” *IEEE Computer*, vol. 40, 2007
- (4) Philipp Wiesner, Ilja Behnke, Dominik Scheinert, Kordian Gontarska, and Lauritz Thamsen. “Let’s wait awhile: how temporal workload shifting can reduce carbon emissions in the cloud.” 22nd International Middleware Conference, 2021
- (5) World Bank, “State and trends of carbon pricing 2022,” tech. rep., 2022
- (6) Noman Bashir, David Irwin, Prashant Shenoy, and Abel Souza. “Sustainable computing - without the hot air.” 1st Workshop on Sustainable Computer Systems Design and Implementation (HotCarbon), 2022
- (7) Ana Radovanovic, et al. “Carbon-Aware Computing for Datacenters”. *IEEE Transactions on Power Systems*, 2022
- (8) Jesse Dodge, et al. “Measuring the carbon intensity of AI in cloud instances.” ACM Conference on Fairness, Accountability, and Transparency (FACCT), 2022
- (9) Shaohong Li, et al. “Thunderbolt: throughput-optimized, quality-of-service-aware power capping at scale.” USENIX Symposium on Operating Systems Design and Implementation (OSDI), 2020
- (10) Liuzixuan Lin, Victor M. Zavala, and Andrew Chien. “Evaluating coupling models for cloud datacenters and power grids.” ACM International Conference on Future Energy Systems (e-Energy), 2021
- (11) Abel Souza, et al. “Ecovisor: a virtual energy system for carbon-efficient applications.” ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), 2023
- (12) Philipp Wiesner, Ramin Khalili, Dennis Grinwald, Pratik Agrawal, Lauritz Thamsen, and Odej Kao. “FedZero: leveraging renewable excess energy in federated learning.” ACM International Conference on Future Energy Systems (e-Energy), 2024
- (13) Philipp Wiesner, Dominik Scheinert, Thorsten Wittkopp, Lauritz Thamsen, and Odej Kao. “Software-in-the-loop simulation for developing and testing carbon-aware applications.” *Software: Practice and Experience*, 2022